DataKitchen

Garbage Data = Garbage AI:

An Open Source Data Quality & Observability For Teams With No Time

Agenda

Perspective: Life is an unproductive mess in data and analytics teams

Fix: principles of DataOps: agile, lean, devops

Problem: AI is awesome and it will make the mess so much worse

Part 1: AI – new use case: data + LLMs to give insight

Part 2: AI – more people creating insight: vibe data engineering

Conclusion

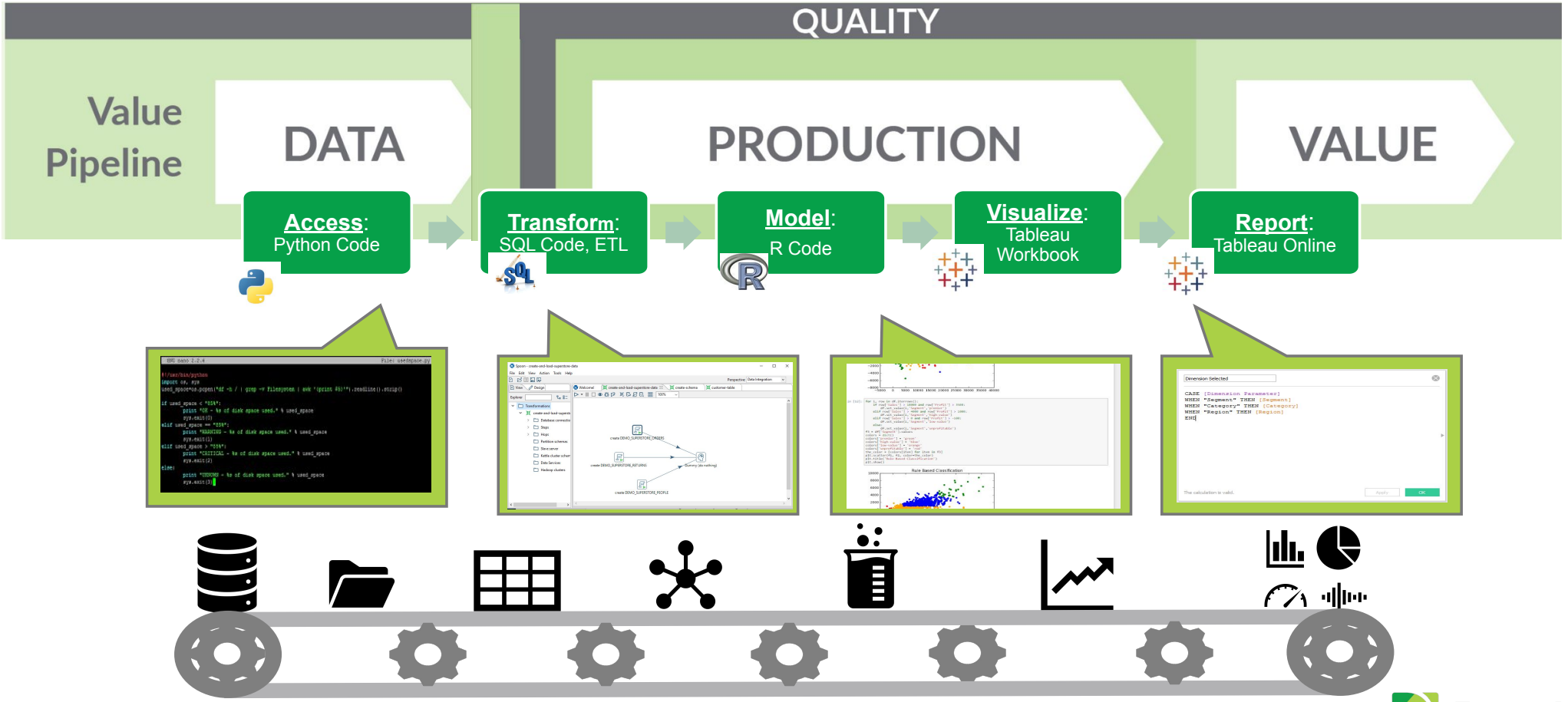
My Background & Focus

- 15 years software & AI: NASA, MIT, startups, Microsoft, then 20 years in data
- I learned three things
  - You get crappy data and shit breaks.
  - Your customers don't know what they want until they see it.
  - You always have too much to do.
- **Data and Analytics have a crisis of productivity**

My Background & Focus

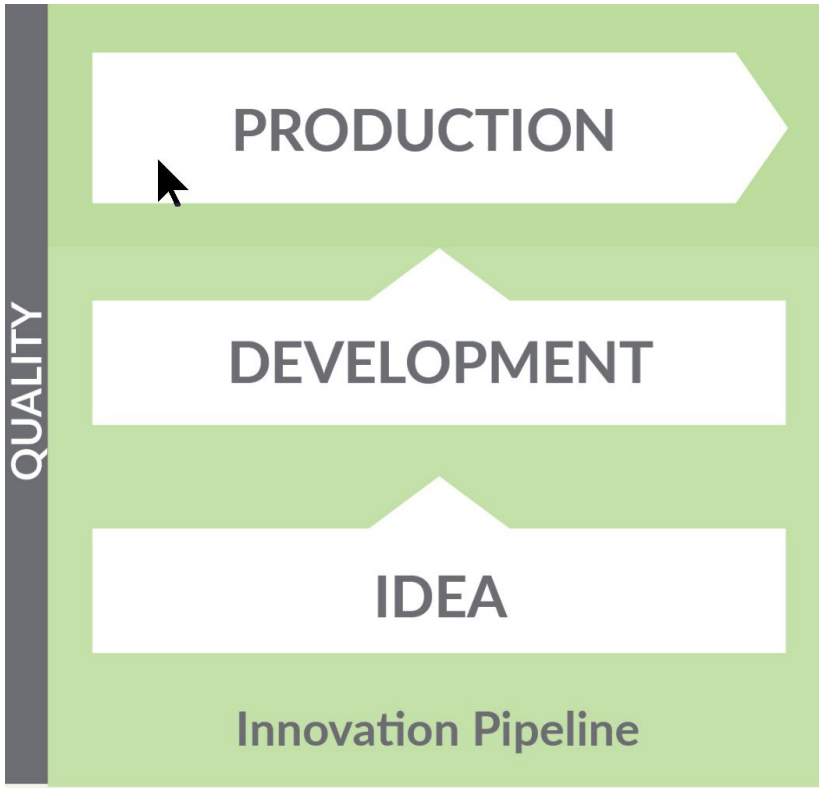
- 15 years software & AI: NASA, MIT, startups, Microsoft, then 20 years in data
- I learned three things
  - You get crappy data and shit breaks.
  - Your customers don't know what they want until they see it.
  - You always have too much to do.
- **Data and Analytics have a crisis of productivity**
- **For the last dozen years we (DataKitchen) are trying to fix this problem: 'DataOps'**
  - We have run a profitable, independent business
    - Data engineering with our software based on DataOps practices
    - Books, trainings, conferences, management consulting, writing, podcasts.
    - Top down software sales to CDOs (ugh)
  - Last few years open source open source data quality & observability to drive DataOps adoption

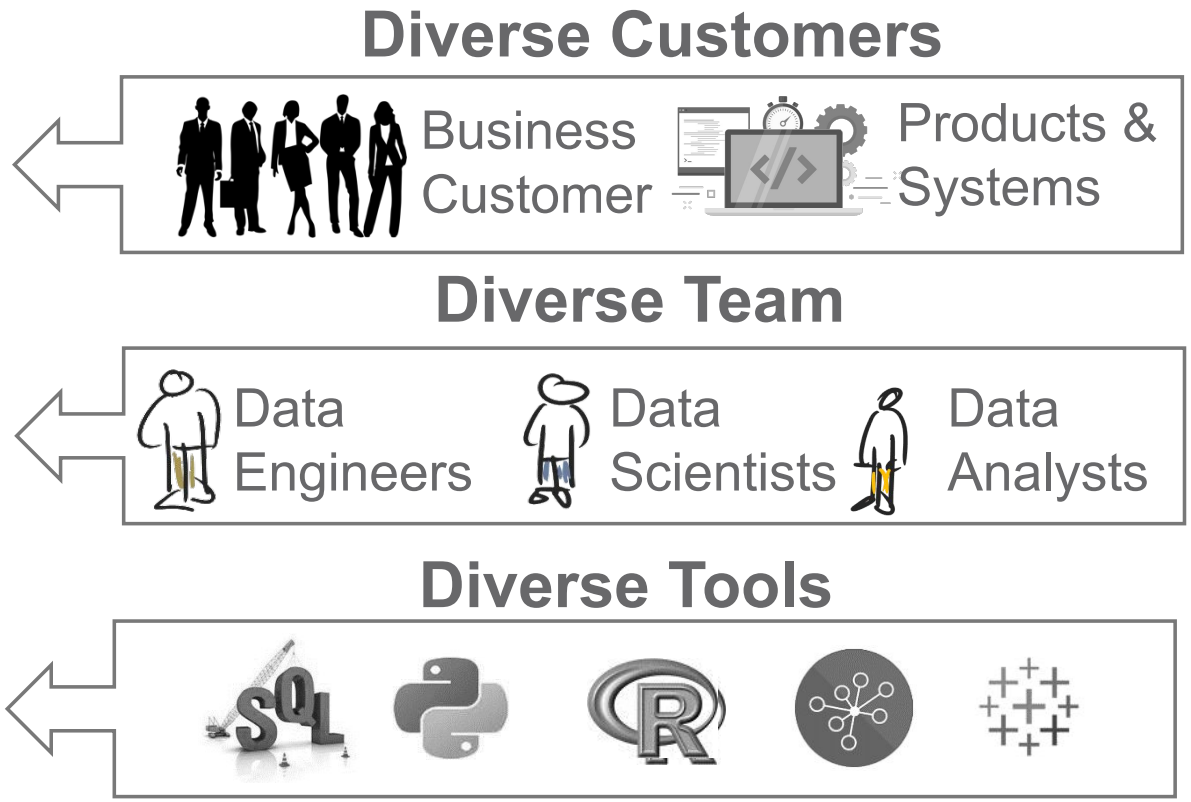
Analytic process are like manufacturing: materials (data) and production outputs (refined data, charts, graphs, model)



DataKitchen

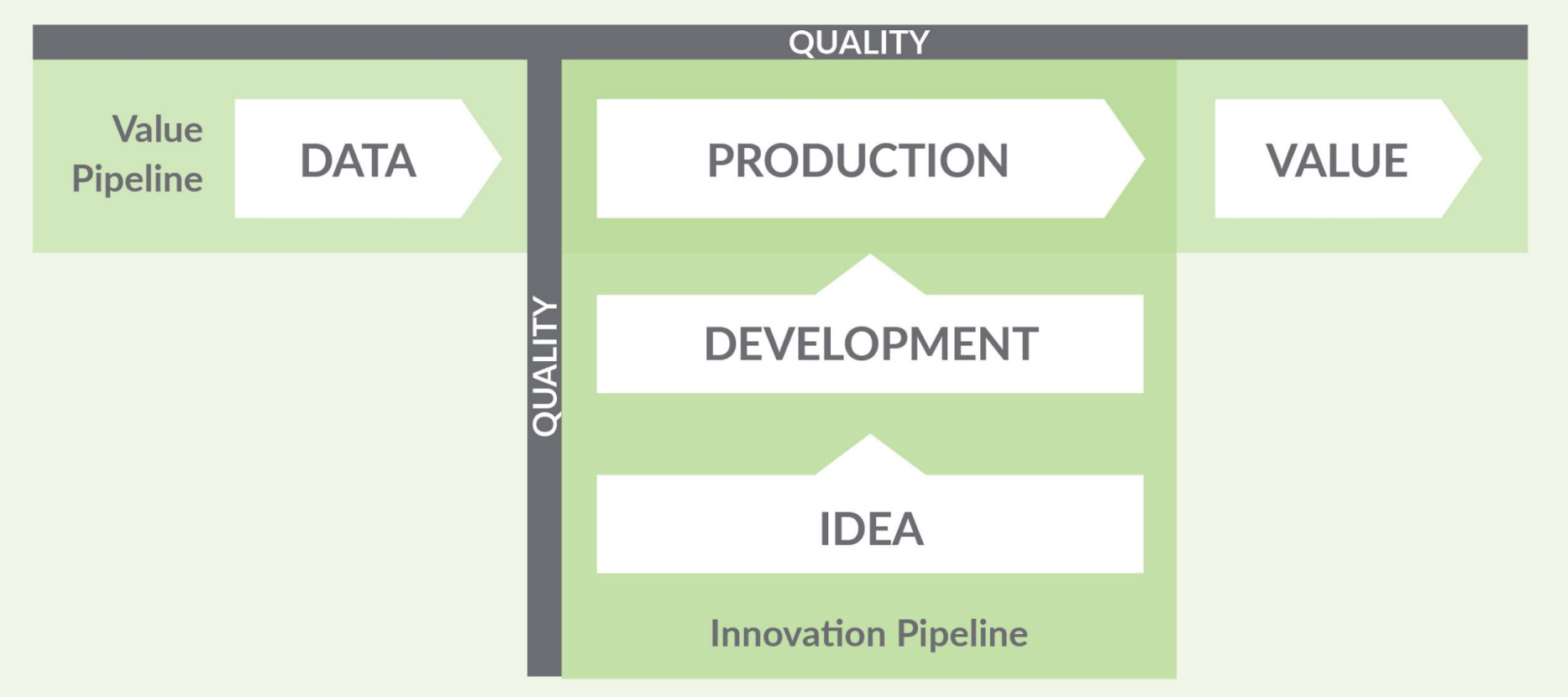
Analytic processes are like software development: code continually move from development to production





DataKitchen

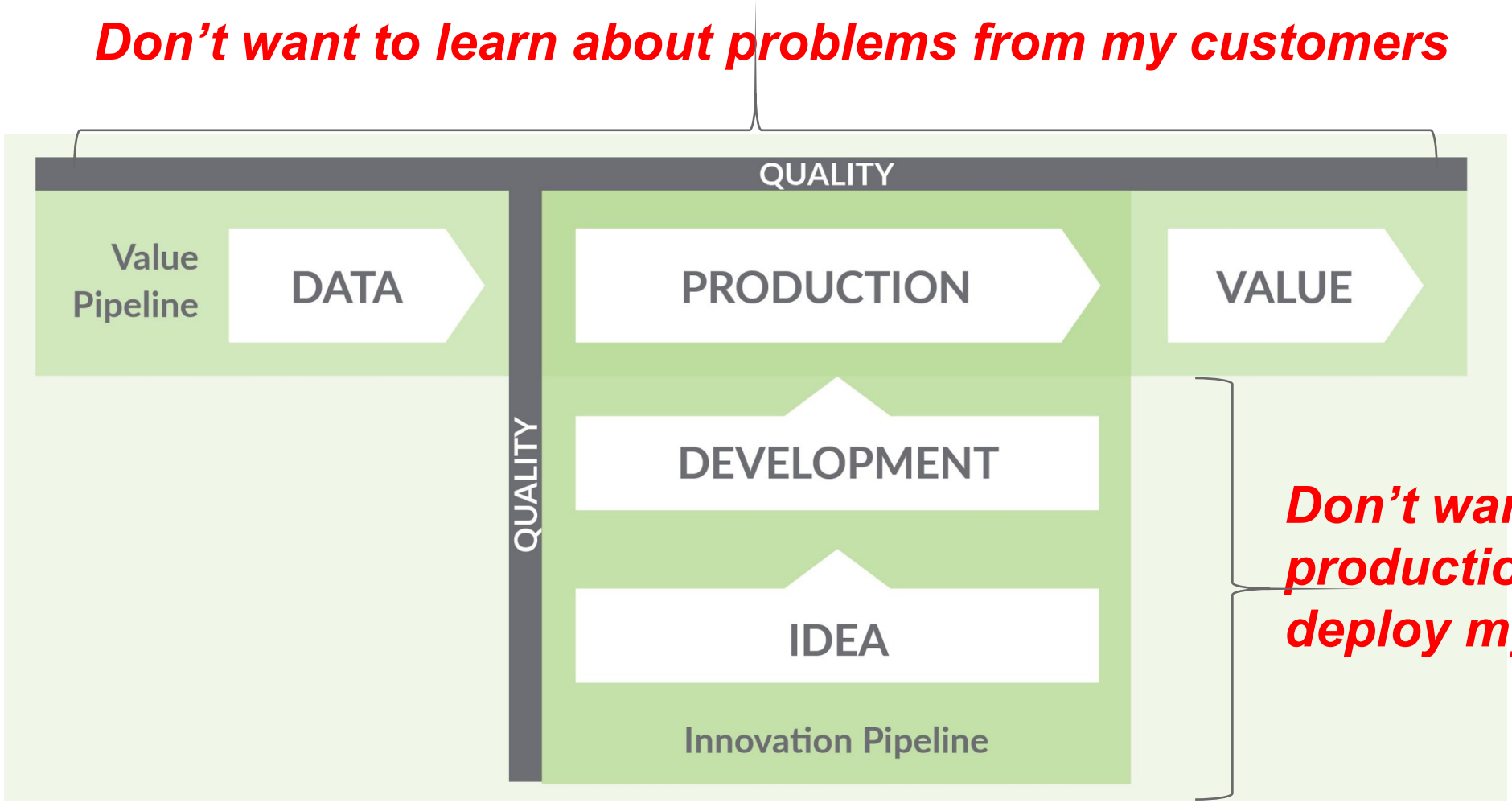
Two Key DataOps Pipelines: Value & Innovation



DataKitchen

What DataOps Helps You Avoid

Don't want to learn about problems from my customers



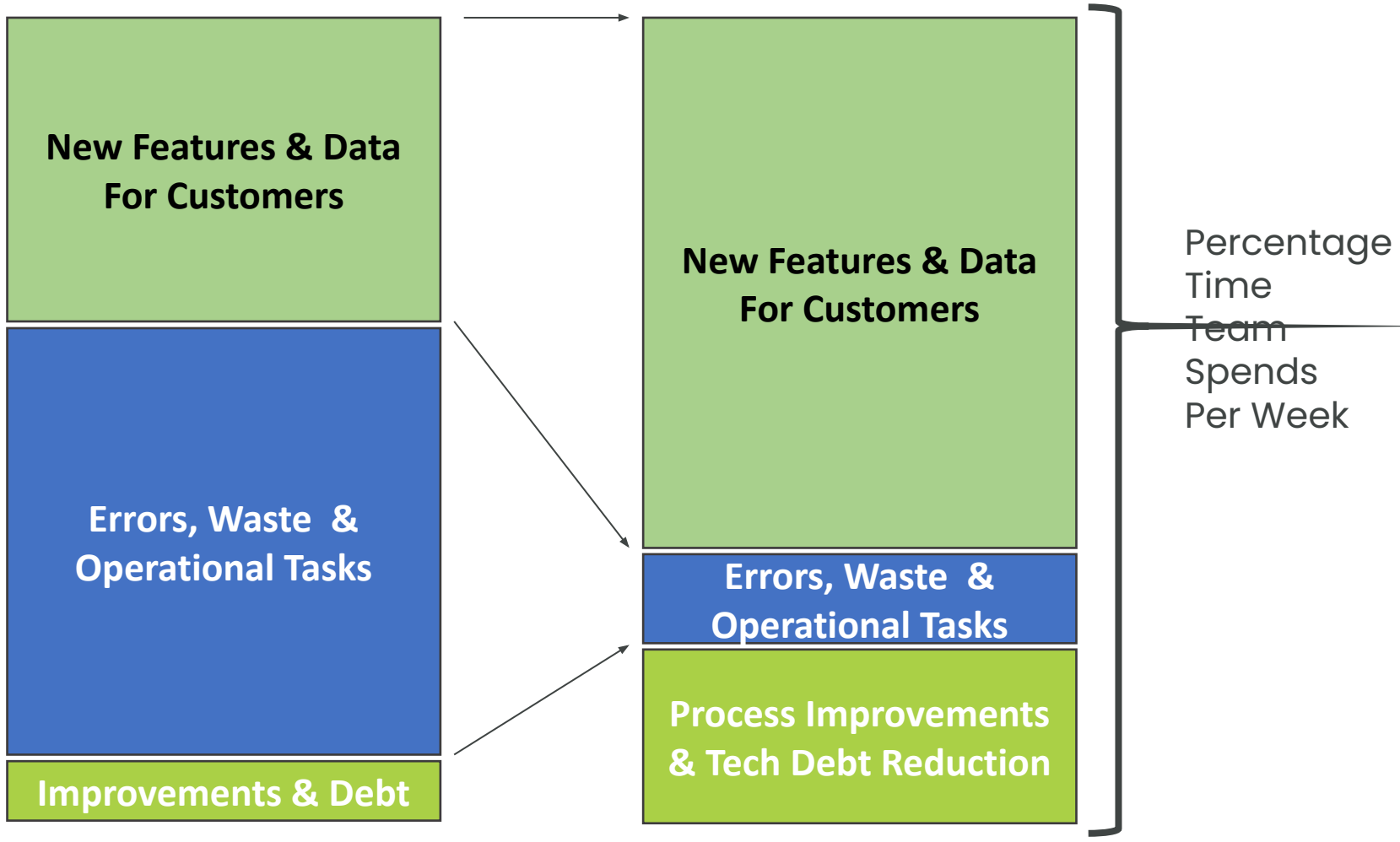
Don't want break production when I deploy my changes

DataKitchen

DataOps Benefit: Improves Productivity

Only 22% of time is spent on innovation;
78% on errors and manual execution.

– Gartner (2020)



DataKitchen

Agenda

Perspective: Life is an unproductive mess in data and analytics, teams

fix: principles of DataOps: agile, lean, devops

Problem: AI is awesome and it will make the mess so much worse

Part 1: AI – new use case: data + LLMs to give insight

Part 2: AI – more people creating insight: vibe data engineering

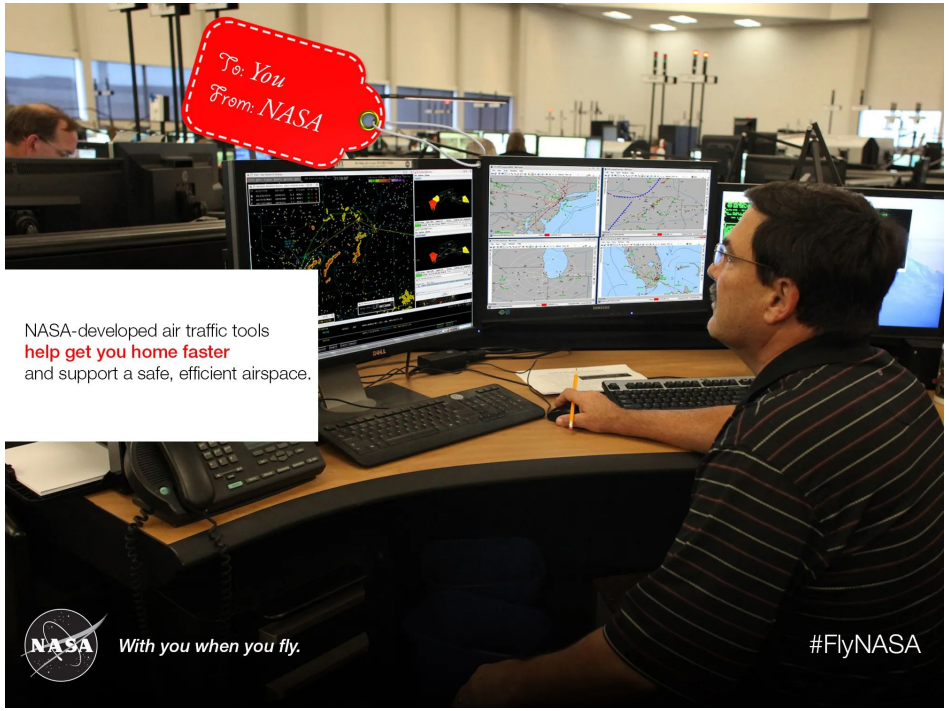
Conclusion

AI Is Invading The Data Team

AI (really Large Language Models or LLMs) reality:

- Your analytic engineers are vibing, coding tables late into the night.
- Your standard reports—people are using Claude to analyze data directly
- Your LLMs are confidently serving up garbage, sometimes.
- Your predictive models—once your pride and joy—are degrading faster than you can debug them.
- More stakeholders are screaming for AI magic.

LLMs (Or Any Model) Will Never Get 100% Perfect Results.



Embrace The Imperfection

- LLMs are somewhere from 10% - 50% **inaccurate.**
- Improving accuracy is getting asymptotically harder.
- LLMs will **ALWAYS** deliver imperfect results

Yet LLMs have a huge potential to **assist** users and make them more productive

AI Is Expanding The World Of Data Analytics

LLMs as an Analytics Interface:

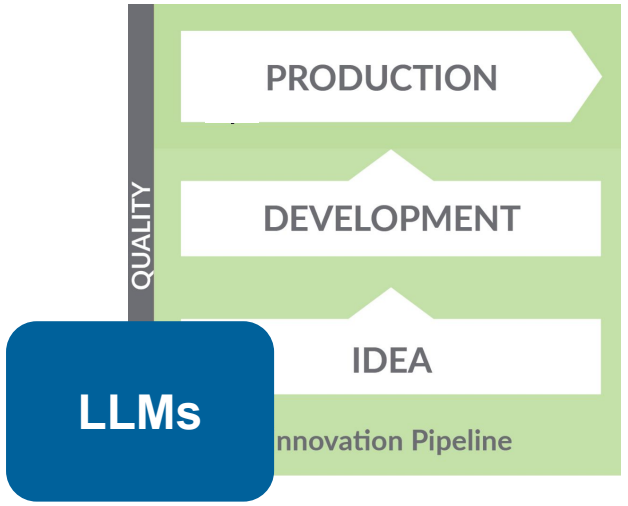
<u>More data</u> being used to make predictions, in new use cases, by utilizing LLMs

DataKitchen

AI Is Expanding The World Of Data Analytics

LLM are expanding the pool of users

<u>More code</u> is being created and put into production by many more people



DataKitchen

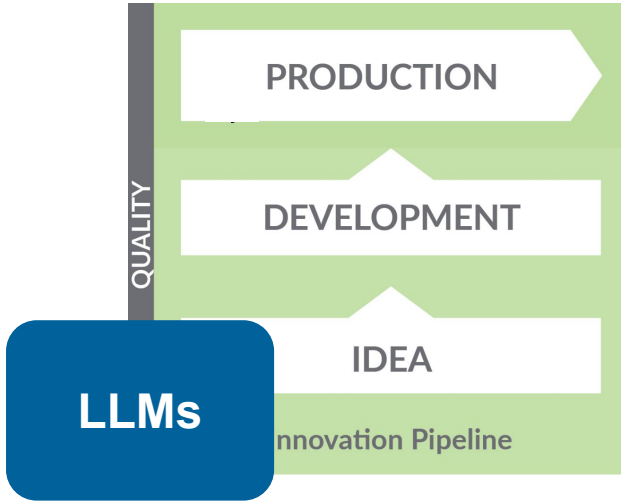
AI Is Expanding The World Of Data Analytics

LLMs as an Analytics Interface:

<u>More data</u> being used to make predictions, in new use cases, by utilizing LLMs

LLM are expanding the pool of users

<u>More code</u> is being created and put into production by many more people



DataKitchen

New Use Case: LLM as an Analytics Interface

Your Data

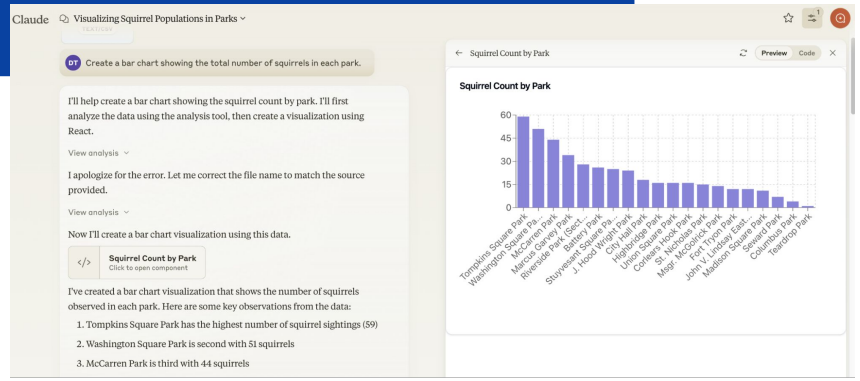
You internal data +
Your internal
metadata / context

Your LLM

ChatGPT, Claude,
Gemini, etc.

Personal Insight

LLM as a ‘conversational
interface’ to data



DataKitchen

LLMs Drive More Data Usage, Faster

Your Data

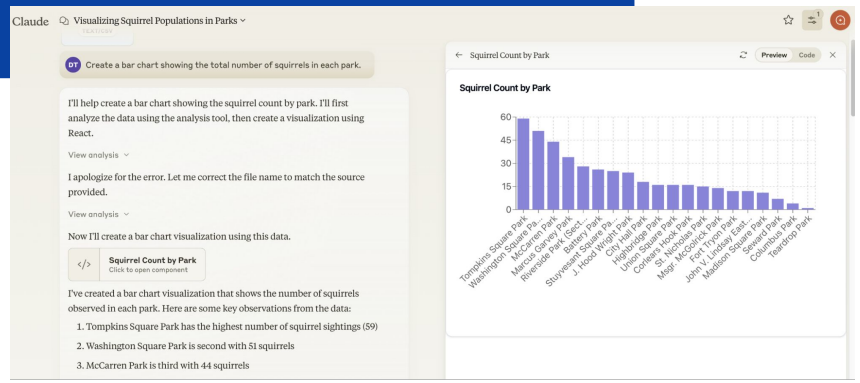
You internal data +
Your internal metadata / context

Your LLM

ChatGPT, Claude, Gemini, etc.

Personal Insight

LLM as a ‘conversational interface’ to data



DataKitchen

Bad Data Compounds LLMs Error

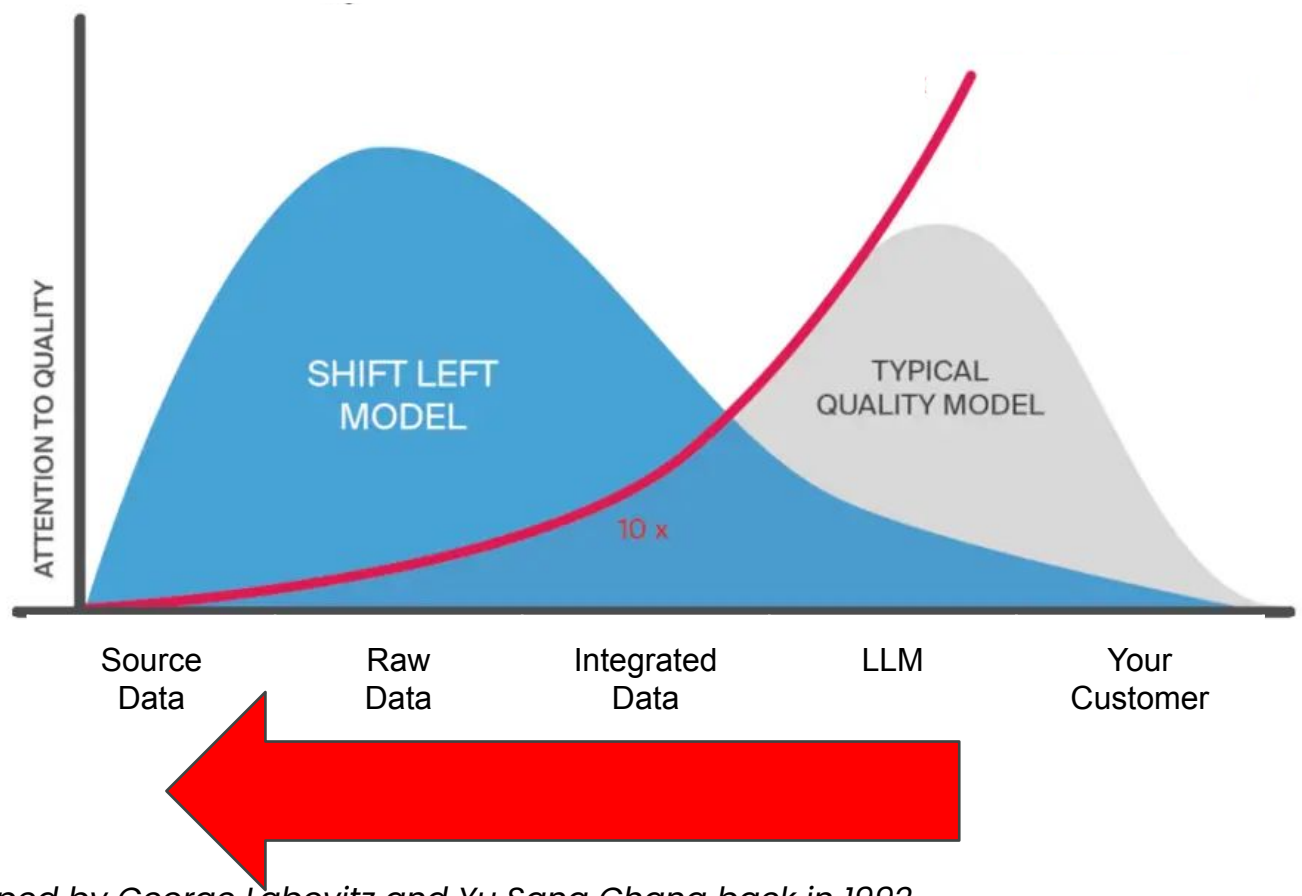
Your Data	Your Model (e.g. LLM)	Your Insight
For Example	For Example	**Expected Value**
Data Is 80% Correct	**Model 80% Accurate**	**64% Accuracy**

Shift Left Data Quality Saves Big Time

It is 10x cheaper to find a problem in data at the source or raw layer.

Apply the 1:10:100 rule*:

- Cost of preventing poor data quality at source is $1 per record
- Cost of remediation after it is created is $10 per record
- Cost of failure (i.e. doing nothing) is $100 per record



*developed by George Labovitz and Yu Sang Chang back in 1992

AI / LLMs Needs Test Coverage

AI are using more data … therefore you need to check data, automatically

… mean you need to improve automated test coverage

In All Levels/Zones In Your Database/Data Lake

- Every Table Should Have Tests
- Every Column In Every Table Should Have Tests
- Every Significant Business/Domain Specific Metric Should Have Tests

Every Tool That Uses Data Should Be Checked For Errors and Timing

DataKitchen

Types of Tests in Test Coverage

What Types Of Automated Tests?

- Every Table Should Have Tests
  - **Consistency Tests: Volume, Freshness, Schema**
- Every Column In Every Table Should Have Tests
  - **Consistency Tests: Volume, Freshness, Schema, Drift**
- Every Significant Business Metric Should Have Tests
  - **Domain Specific: Custom, Domain Specific Testing**

Every Tool That Uses Data Should Be Checked For Errors and Timing

- **Check Logs For Errors**
- **Check Metrics**
- **Check Task Status and Substatus Results**
- **Check Timing and Duration**

DataKitchen

How to Measure Test Coverage

In All Levels/Zones In Your Database/Data Lake

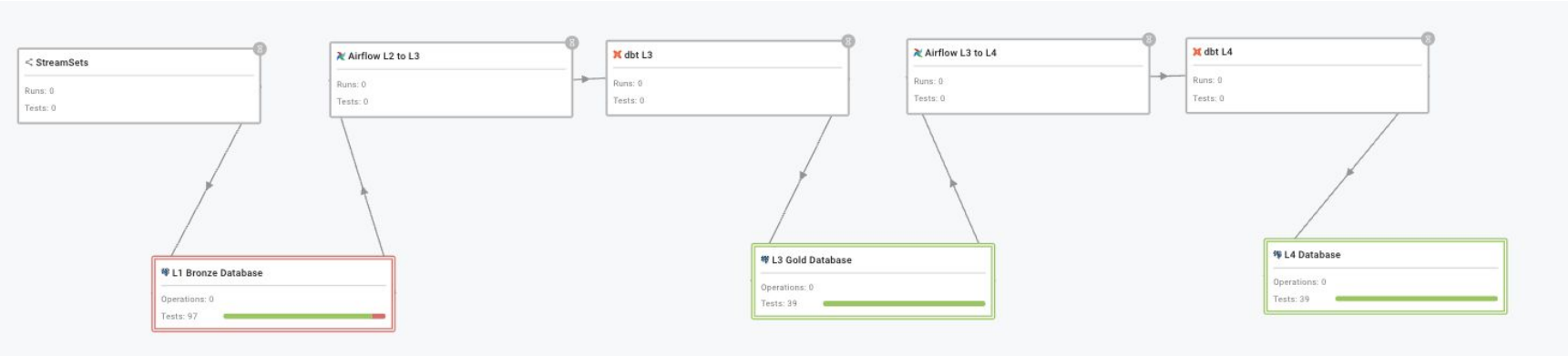
- Every Table Should Have Tests
  - **Minimum 2 Test Per Table**
- Every Column In Every Table Should Have Tests
  - **Minimum 2 Tests Per Column**
- Every Significant Business Metric Should Have Tests
  - **Minimum 1 Custom Test Per Metric**

Every Tool That Uses Data Should Be Checked For Errors and Timing

- **Minimum 1 Check Per Tool:** Logs For Errors
- **Optional**: Check Metrics
- **Minimum 1 Check Per Tool Per Job:** Task Status and Substatus Results
- **Minimum 1 Check Per Tool Per Job:** Timing and Duration

DataKitchen

Example: Test Coverage Counts Medallion



This Medallion Architecture Has **Three Database Levels** (L1, L2, L3), **Three Tools** (Streamsets, Airflow, dbt), And **Five Jobs/Workflows** to Monitor

- *L1 - 100 tables + 10 Columns*: Minimum 200 Table Tests and 2000 Column Tests
- *L2 - 100 tables + 10 Columns*: Minimum 200 Table Tests and 2000 Column Tests
- *L3 - 10 tables + 30 Columns*: Minimum 20 Table Tests and 600 Column Tests, Plus a handful of Domain Specific Business Metric Tests
- *3 Tools*: 3 Checks For Errors
- *6 Jobs*: 12 Checks For Problems

DataKitchen

Writing Data Test Manually Does Not Scale

An example

- With TestGen, a junior operator can generate 2,500 tests with **two steps** (profile, generate tests)
- It would take a trained Data Engineer **7.2 months** to achieve the same results – with no time for meetings, breaks, or vacations

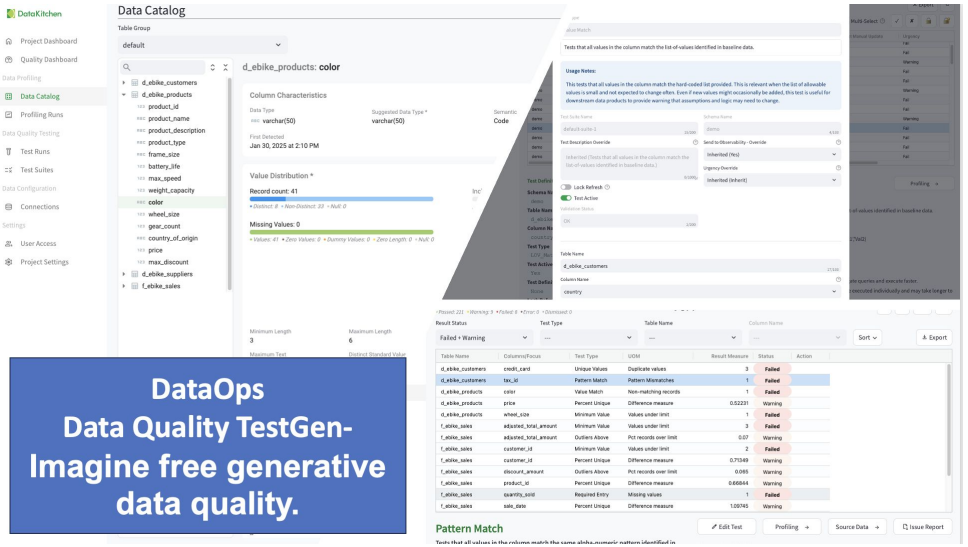
Number of Tests	**2,500**	
Number of Tables	20	
Number of Columns	1,000	
Minutes per Test	30	
Hours	1,250	
Days	156	8 hours/day
Weeks	31	5 days/week
Months	**7.2**	4.35 weeks/month
TestGen	**2 steps**	

How long does it take to write one data quality test?
Do you engineers know what tests to write?
Need AI to fight AI

DataKitchen

DataKitchen Data Quality TestGen Software: 80% of the data tests you need automatically

- **One-Click Data Quality –** Instantly generate and run automated tests.
- **Data Profiling –** 51 column-level insights.
- **In-Database Execution –** Fast, secure testing in your own database.
- **120+ AI-Generated & Custom Tests –** Comprehensive data validation coverage.
- **Anomaly Detection –** Automatic alerts for freshness, volume, schema, and drift.
- **Data Catalog –** Unified view of metadata, hygiene issues, PII, and test results.
- **Quality Scoring Dashboards –** Custom scorecards with drill-downs to drive improvement.



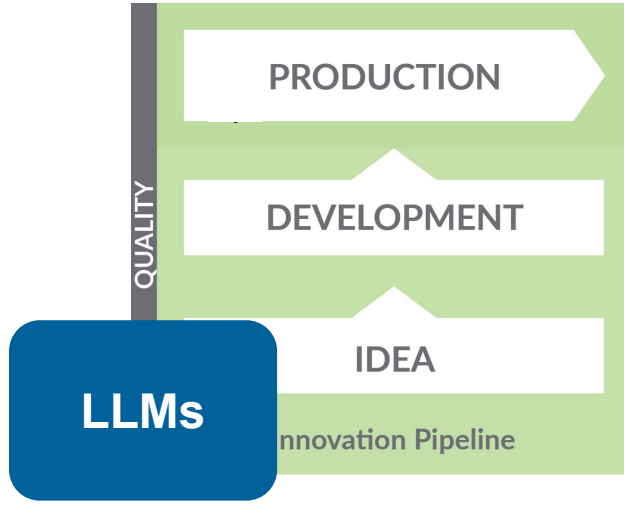
AI Is Expanding The World Of Data Analytics

LLMs as an Analytics Interface:

More data being used to make predictions, in new used cases, by utilizing LLMs

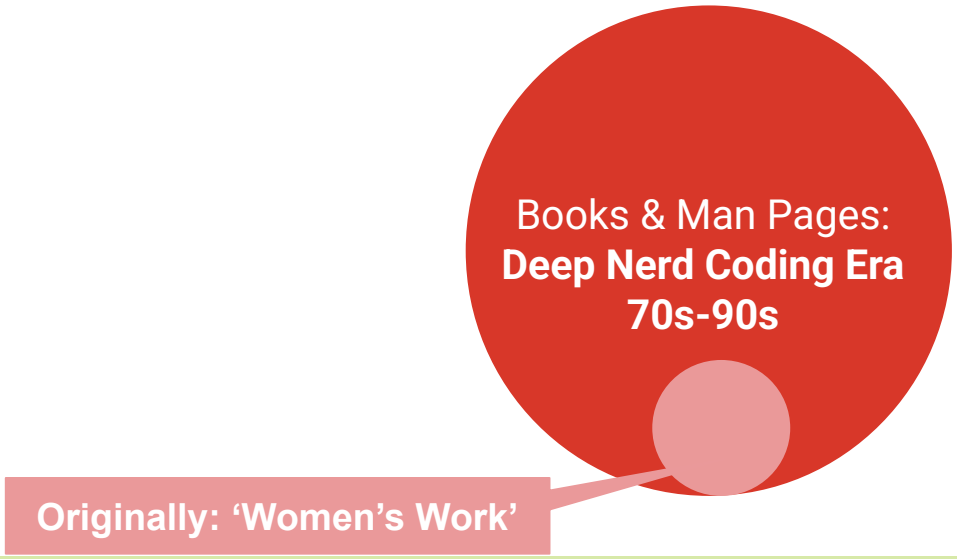
LLM are expanding the pool of users

More code is being created and put into production by many more people

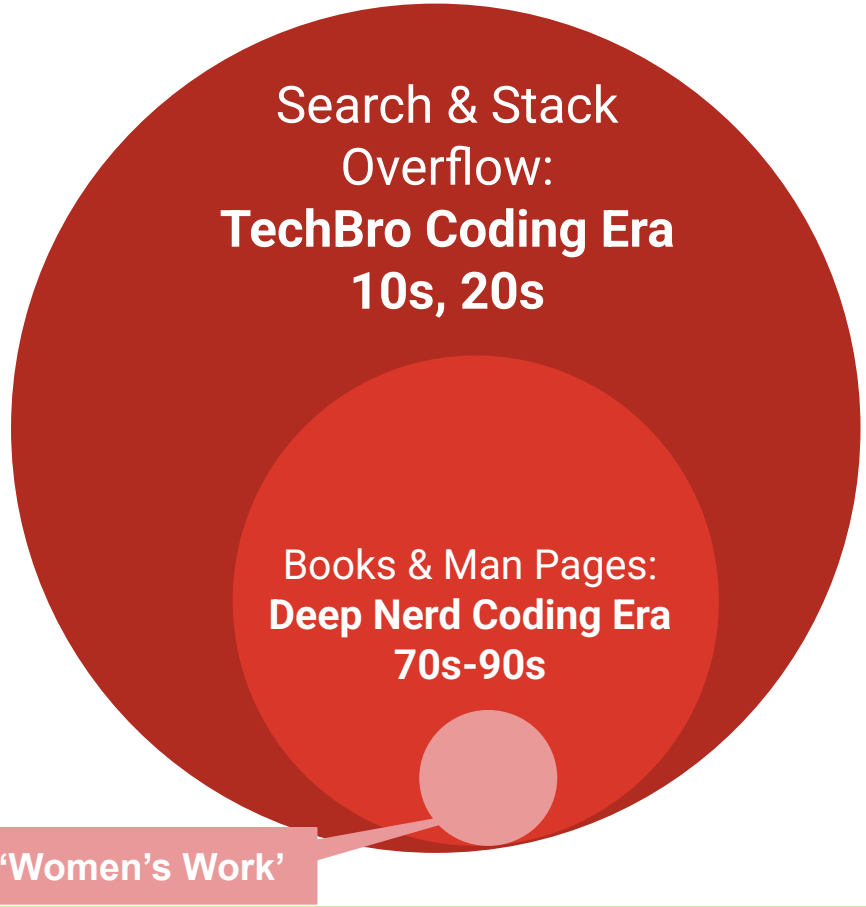


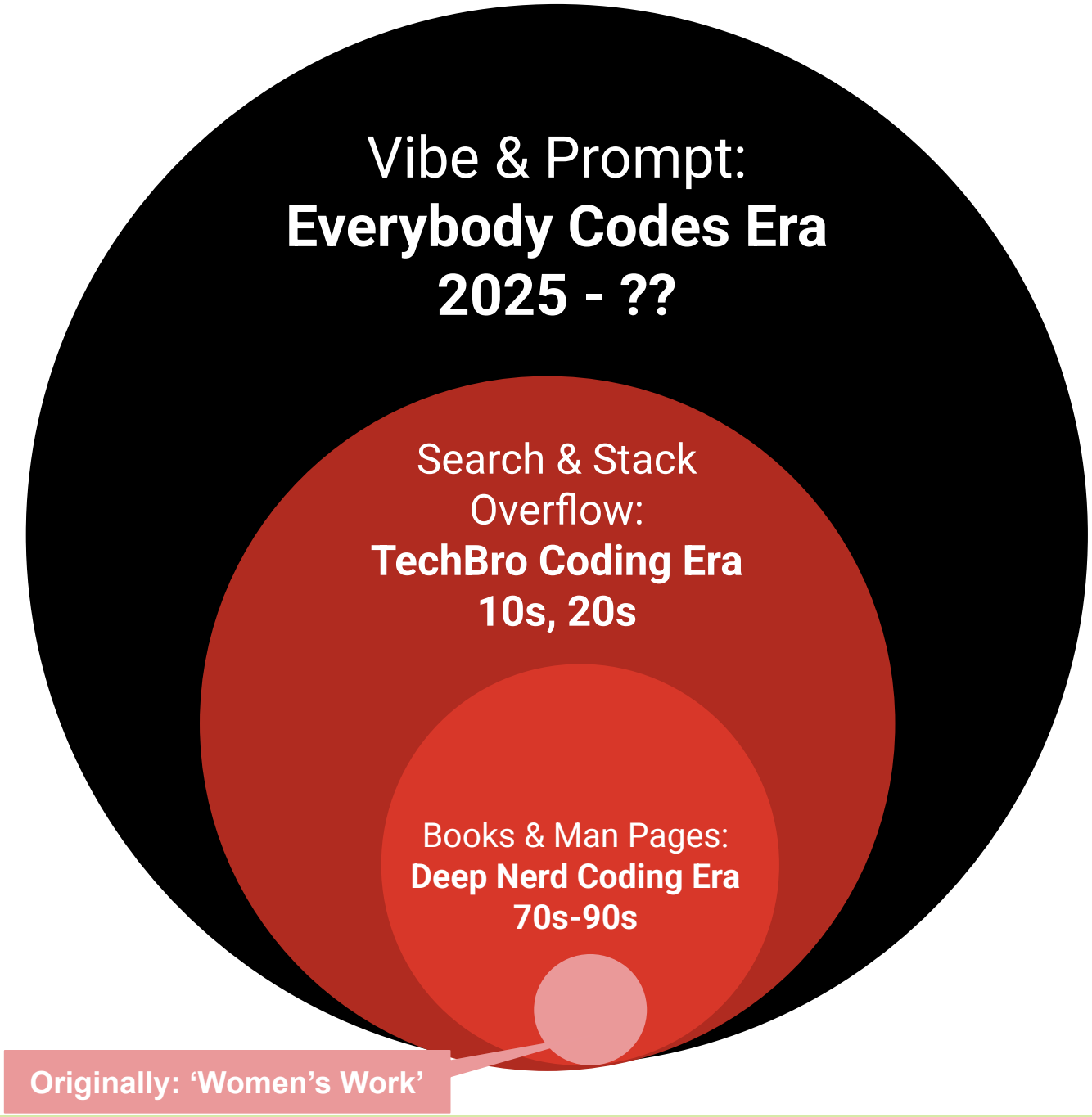
DataKitchen

Coding: A New Era of Data Teams



Coding: A New Era of Data Teams





Coding: A New Era of Data Teams

LLMs increase the production of Analytics

More people will be building

- new tables
- new reports
- new models
- new use cases
- and lots of SQL

It's a vibe coding extravaganza!

AI Is Expanding the World Of Data

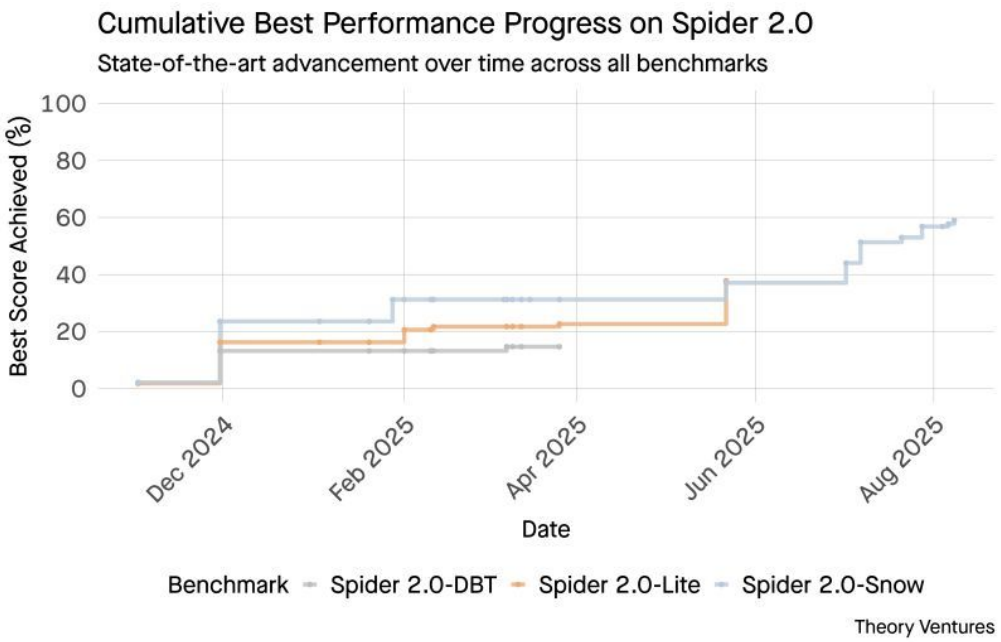
More code being created in data analytics systems, by more types of people, due to LLMs. For example:

- Using Claude Desktop to help write SQL
- Using the 'AI Assistant' in your favorite ETL or orchestration tool help write YAML configuration
- Using Claude Code to write some Python model code.

You boss, your customer are all doing these kinds of actions today

There are more an more people coding … and more coming!

LLMs Struggle With Writing SQL



- GPT-5 excels at math but struggles with databases
- Spider 2.0 benchmarks expose the gap
- Performance remains poor across all variants:
  - Spider 2.0-Snow peaks at 59.05% accuracy, Spider 2.0-DBT tops out at 39.71%
  - 56 submissions from 12 model families since November 2024.
- Business context is the real challenge
- Human judgment remains irreplaceable

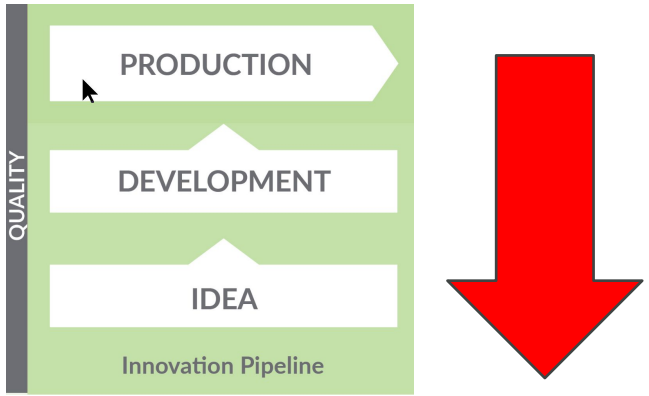
Context Helps LLM Improve Accuracy

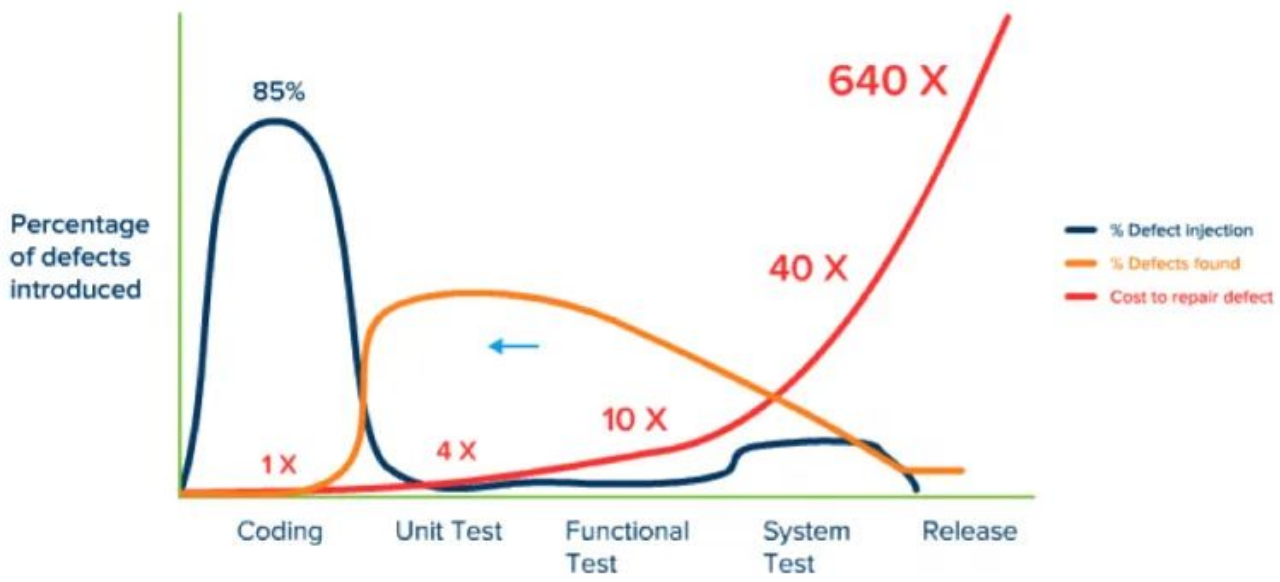
- AI generates code that looks right but fails
- For example, it produces syntactically correct SQL but doesn't understand your actual data, creating solutions that break in practice.
- Why?
  - Real data is messy
  - Data engineering needs context, not just code
  - "Data Context" formalizes hidden knowledge
  - Context enables AI to reason better
    - Data Profiling, Data Catalog, Data Quality Testa Results, Etc.

Context Is Not A Miracle Cure: Expect More Code, Worse Quality

Shift Down Test Coverage Saves Costs

The earlier your tests find a problem in development, the lower the cost and hassle

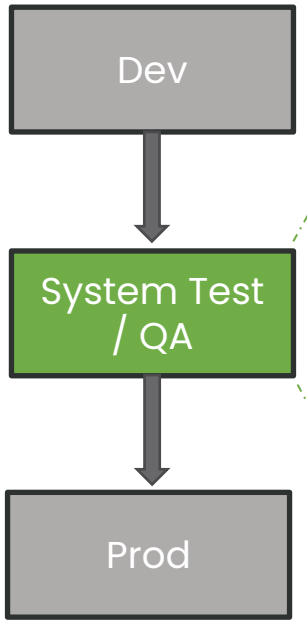




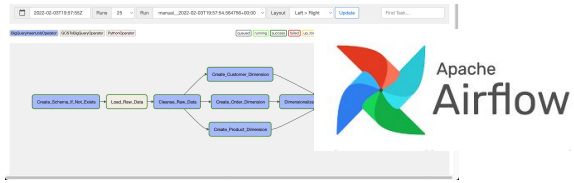
Jones, Capers. *Applied Software Measurement: Global Analysis of Productivity and Quality.*

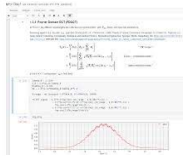
DataKitchen

Regression Needs Full Coverage of Tools & Data

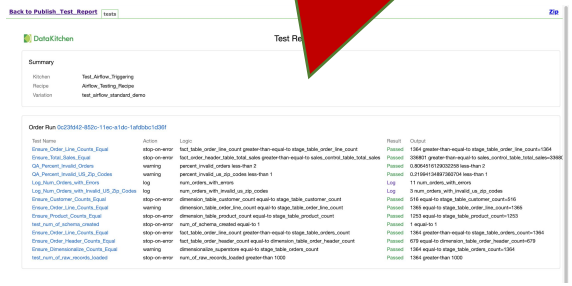


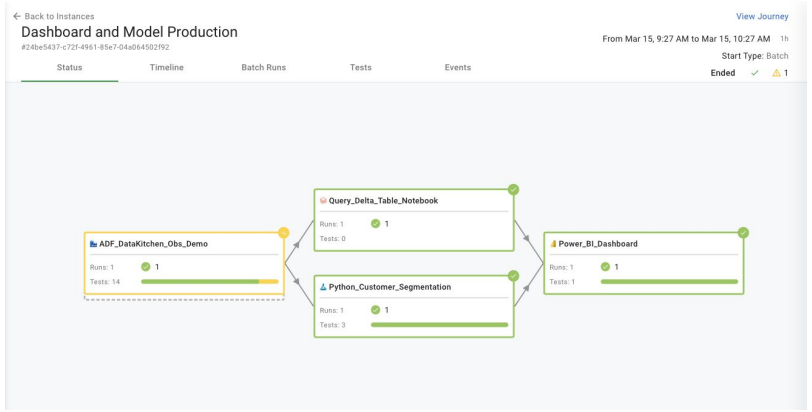
A CI /CD Process (Jenkins)



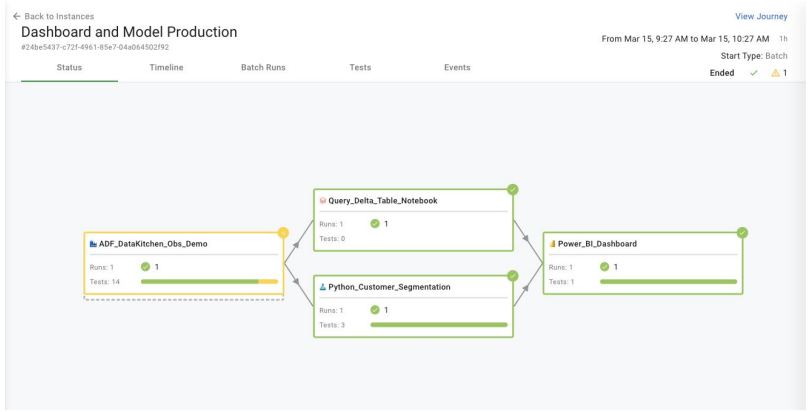


Find Code Problems in Development





DataOps Observability: Data Observability Tool



DataOps Observability:

- Open Source, Full Featured, Data Observability Tool. All Features, One User
- Enterprise Version Starts At $100 Per User Per Month
- Extensive Connectors

It Does Five Tasks:

1. Single Pane Of Glass Across Your Entire Data Production State
2. Process Lineage Graph – Data Journey
3. Collects Logs, Metrics, Runs Status, Schedules, And Test Result
4. Production Dashboard And Alerts
5. Enables Full Regression Testing And Production Monitors and Andons

DataKitchen

Agenda

Perspective: Life is an unproductive mess in data and analytics, teams

fix: principles of DataOps: agile, lean, devops

Problem: AI is awesome and it will make the mess so much worse

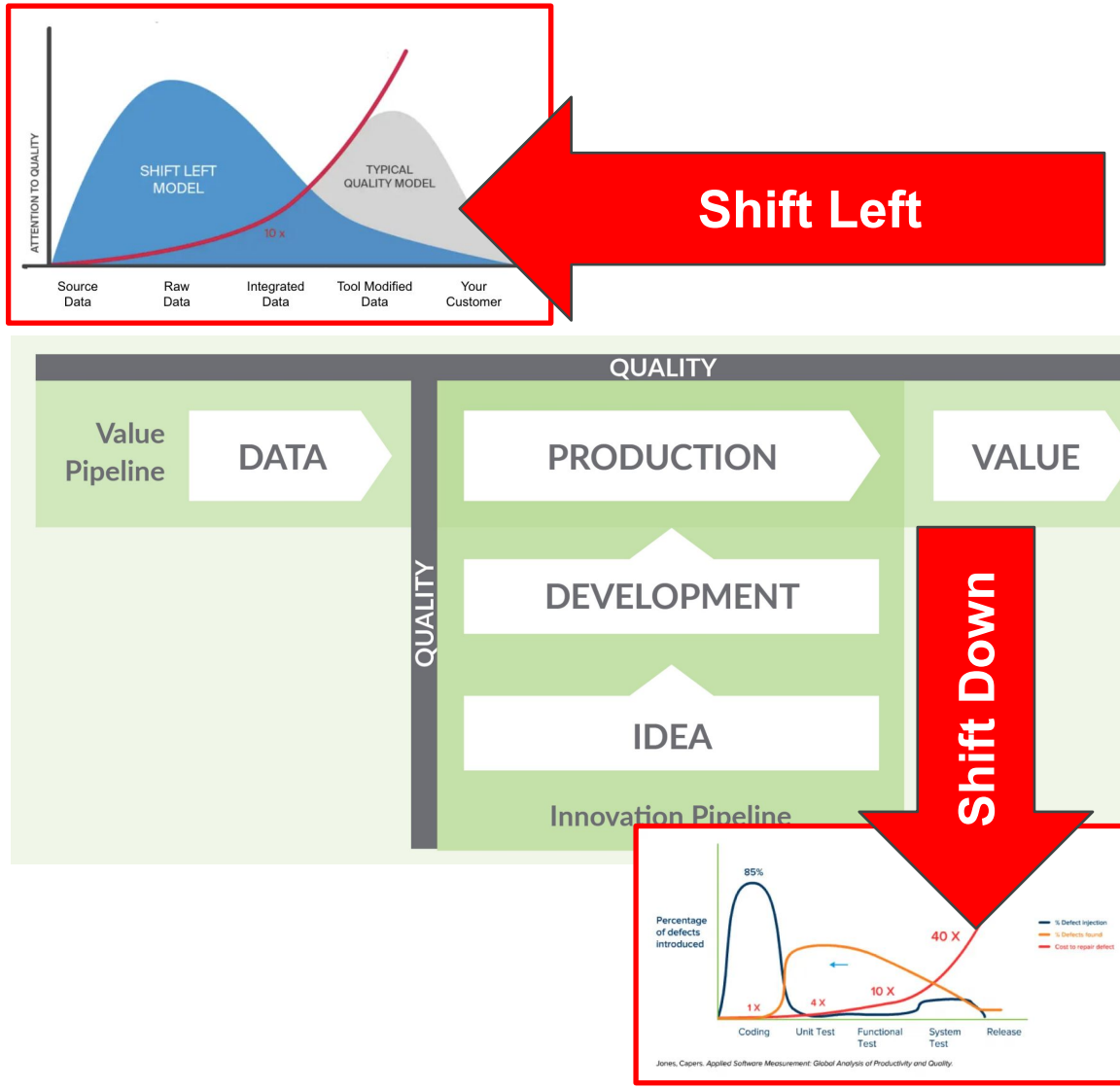
Part 1: AI – new use case: data + LLMs to give insight

Part 2: AI – more people creating insight: vibe data engineering

Conclusion

Data Quality and Observability for Small Data Teams

- Low Code
- AI to fight AI
- Full Featured Open Source
- White Glove User Experience
- Enterprise: Fixed-Fee Pricing. Unlimited Usage.



Learn More About DataOps & Data Observability

Install Open Source TestGen
https://info.datakitchen.io/testgen

Install Open Source DataOps Observability
https://docs.datakitchen.io/articles/#!open-source-data-observability/install-data-observability-products-open-source

Sign The DataOps Manifesto
http://dataopsmanifesto.org

Free DataOps Cookbook
https://datakitchen.io/the-dataops-cookbook/

Free DataOps Certification
https://info.datakitchen.io/training-certification-dataops-fundamentals

Free Data Quality & Observability Certification
https://info.datakitchen.io/data-observability-and-data-quality-testing-certification

DataKitchen